Manuscript prepared for Geosci. Model Dev.
with version 4.2 of the LaTeX class copernicus.cls.
Date: 28 November 2011

Land surface Verification Toolkit (LVT) - A generalized framework for land surface model evaluation

Sujay. V. Kumar[1,2], Christa. D. Peters-Lidard[2], Joseph Santanello[2], Ken Harrison[2,3], Yuqiong Liu[2,3], and Michael Shaw[1,2,4]

[1]Science Applications International Corporation, Beltsville, MD
[2]Hydrological Sciences Laboratory, NASA Goddard Space Flight Center, Greenbelt, MD
[3]Earth System Science Interdisciplinary Center, College Park, MD
[4]Air Force Weather Agency, Offutt, NE

Correspondence to: Sujay V. Kumar
(Sujay.V.Kumar@nasa.gov)

Abstract. Model evaluation and verification are key in improving the usage and applicability of simulation models for real-world applications. In this article, the development and capabilities of a formal system for land surface model evaluation called the Land surface Verification Toolkit (LVT) is described. LVT is designed to provide an integrated environment for systematic land model evaluation and facilitates a range of verification approaches and analysis capabilities. LVT operates across multiple temporal and spatial scales and employs a large suite of in-situ, remotely sensed and other model and reanalysis datasets in their native formats. In addition to the traditional accuracy-based measures, LVT also includes uncertainty and ensemble diagnostics, information theory measures, spatial similarity metrics and scale decomposition techniques that provide novel ways for performing diagnostic model evaluations. Though LVT was originally designed to support the land surface modeling and data assimilation framework known as the Land Information System (LIS), it also supports hydrological data products from other, non-LIS environments. In addition, the analysis of diagnostics from various computational subsystems of LIS including data assimilation, optimization and uncertainty estimation are supported within LVT. Together, LIS and LVT provide a robust end-to-end environment for enabling the concepts of model data fusion for hydrological applications. The evolving capabilities of LVT framework are expected to facilitate rapid model evaluation efforts and aid the definition and refinement of formal evaluation procedures for the land surface modeling community.

1 Introduction

Verification and evaluation are essential processes in the development and application of simulation models. Land surface models (LSMs) are one such class of simulation models specifically designed to represent the terrestrial water, energy and biogeochemical processes. LSMs generate estimates of terrestrial biosphere exchanges by solving governing equations of soil-vegetation-snowpack medium, and can be run in either offline mode or coupled to an atmospheric model. An accurate representation of land surface processes is therefore critical for improving models of the boundary layer and land-atmosphere coupling as well as real world applications such as ecosystem modeling, agricultural forecasting and water resources prediction and management (NRC (1996)). The process of systematic evaluation and verification helps in the characterization of accuracy and uncertainty in the model predictions, which can then be used as a benchmark for future model enhancements. Further, quantitative measures of the fidelity of model simulations are essential for improving the usage and acceptability of LSM forecasts for real-world applications.

The Global Energy and Water Cycle Experiment (GEWEX) Global Land Atmosphere System Study (GLASS) has identified that a general benchmarking framework capable of capturing useful modes of variability of LSMs through a range of performance metrics is necessary for further advancing the performance and predictability of the models (van den Hurk et al. (2011)). In their recommendation of the priorities for hydrologic research, Entekhabi et al. (1999) emphasize the need for defining formal evaluation procedures to improve the "observability" of many LSM processes. For e.g., soil moisture in most LSMs represents an index of the moisture state (Koster et al. (2009)) and the estimates from different models vary significantly even when forced with the same meteorology (Dirmeyer et al. (2006)). Further, the soil profile representations in LSMs and assumptions about parameters such as soil hydraulic properties vary significantly across models. As a result, direct comparison of soil moisture estimates from these models against in-situ and remote sensing measurements becomes difficult. Given that a large suite of application models require soil moisture estimates as inputs (e.g. weather and climate forecasting (Fennessey and Shukla (1999); Koster et al. (2004)), agricultural models (Rosenzweig et al. (2002)), ecosystem models (Friend and Kiang (2005))), it is important for the LSMs to generate observable estimates of soil moisture to avoid potential misinterpretations and incorrect usages. The development of a formal, systematic environment for model evaluation will help in bridging the gaps between the model and observations and in improving the observability of LSM outputs.

Model performance is typically improved by either enhancing the conceptual representations of processes (i.e., model physics) or by employing computational techniques (e.g., data assimilation, optimization, uncertainty algorithms, fuzzy logic) to augment model simulations. These computational techniques provide the tools to exploit the information content in the observational data for improving model predictions. The concept of "model data fusion" (MDF; Raupach et al. (2005); Williams et al. (2009)) has been used to describe the paradigm of combining the information from

models and available datasets. The key aspect of the MDF philosophy consists of using information from data to help the formulation, characterization and evaluation of models in a structured manner. The results of the evaluation step are then used to revise and improve model formulation and subsequent development. As part of the new structure formulated in 2009, the GLASS community has identified Benchmarking and MDF as two of its three core themes for research going forward. Here we describe the development of a formal evaluation system for land surface models that addresses both these themes identified by the GLASS community. The evaluation framework is designed to supplement an existing modeling system, to enable end-to-end formulations of the MDF paradigm.

As described in Kumar et al. (2006), Peters-Lidard et al. (2007) and Kumar et al. (2008a), the NASA Land Information System (LIS) is a flexible land surface modeling framework that has been developed with the goal of integrating satellite- and ground-based observational data products and advanced land surface modeling techniques to produce optimal fields of land surface states and fluxes. The LIS infrastructure is designed as a land surface modeling and hydrological data assimilation system that generates estimates of water and energy states (e.g. soil moisture, snow) and fluxes (e.g. evaporation, transpiration, runoff) over a range of spatial (as finely resolved as 1km or finer) and temporal (up to 1 hour and finer) resolutions. LIS operates several community land surface models and supports their application over global, regional or point domains. LIS is designed with advanced software engineering principles and provides a flexible, extensible framework for the inclusion of models, computational tools and datasets.

As a land surface modeling component for earth system models, LIS has also been coupled to atmospheric models such as the Weather Research and Forecasting (WRF) model (Kumar et al. (2007); Santanello et al. (2009)). LIS includes a comprehensive data assimilation subsystem (Kumar et al. (2008b)) that enables the incorporation of several observational and satellite data sources for assimilation, in an interoperable manner. Additional computational tools to assist the utilization of data include parameter estimation and optimization (Santanello et al. (2007); Peters-Lidard et al. (2008); Kumar et al. (2011)) and uncertainty modeling (Harrison et al. (2011)) subsystems. The uncertainty modeling components in LIS enable the explicit characterization of different sources of uncertainty in modeling using Bayesian inference techniques. In summary, LIS provides several key components of the MDF paradigm, including a suite of LSMs and computational tools such as data assimilation, optimization and uncertainty estimation.

In this article, we describe the development of a formal system for land surface model evaluation called the Land surface Verification Toolkit (LVT), designed to enable the systematic evaluation and intercomparison of various terrestrial hydrological datasets. LVT not only supports the diagnostic evaluation of the land model simulations from LIS and other land surface modeling systems, but also provides the capabilities for the analysis of outputs from various LIS subsystems such as data assimilation, optimization, uncertainty estimation, radiative transfer and emission models, and application models. A large suite of in-situ, remotely-sensed and other model and reanalysis datasets

are supported in LVT, which captures a wide range of land surface and terrestrial hydrologic regimes across the globe. In addition, a wide range of analysis metrics and procedures are supported in LVT to facilitate a comprehensive evaluation of hydrological datasets. Figure 1 presents a schematic of the key functions of LVT and its interconnections with LIS and the observational datasets. The following sections describe the capabilities of LVT in detail.

Together, LIS and LVT encompass a comprehensive set of computational tools for fully enabling the MDF concept. The capabilities in LIS enable the estimation of model parameters with the use of the optimization subsystem and state estimation with the use of the data assimilation subsystem. The uncertainty estimation tools enable the characterization of various sources of input uncertainty and their impacts on model prediction uncertainty. By providing the tools for model testing and diagnostic evaluation, LVT completes the requisite components of the MDF paradigm.

This article is structured as follows: Section 2 provides a review of the land model evaluation and verification efforts. This is followed by the description of LVT design (Section 3) and features (Section 4). A number of examples are presented in Section 5 that demonstrate how the LVT capabilities enable end-to-end MDF experiments.

2 Background

There have been a number of efforts to document and standardize land surface model evaluation. The model process development studies are typically focused on evaluating the model performance at point or local scales (e.g., Henderson-Sellers et al. (1995); Chen et al. (1996); Pitman and Henderson-Sellers (1998); Koren et al. (1999); Blyth et al. (2010); Barlage et al. (2010); Niu et al. (2011)). Though they are instrumental in benchmarking the improvements to model physics, these reported enhancements do not necessarily translate to broader spatial scales. Blyth et al. (2011) stresses that the model evaluations must be performed separately at the scales of interest, to guarantee transferability of model processes to different scales.

There have been several community-wide efforts such as the Global Soil Wetness Project (GSWP; Dirmeyer et al. (2006)), African Monsoon Multidisciplinary Analysis (AMMA) Land surface Model Intercomparison Project (ALMIP; de Rosnay et al. (2006)) and Carbon-LAnd Model Intercomparison Project (C-LAMP; Randerson et al. (2009)) that were focused on evaluating and intercomparing a suite of land surface models when forced with a common suite of inputs. These studies documented the systematic improvements in land surface model development and provided benchmarks for the simulation of continental scale water and energy budgets. Similar multi-model efforts include the North american Land Data Assimilation System (NLDAS; Mitchell et al. (2004)) and the Global Land Data Assimilation System (GLDAS; Rodell et al. (2004b)) projects, which generate land surface model outputs in near real-time, forced with observation-based meteorology. A detailed evaluation of the NLDAS model products against available observations were conducted during phase-I

and II of the project (Robock et al. (2003); Sheffield et al. (2003); Pan et al. (2003); Lohmann et al. (2004); Mo et al. (2011); Xia et al. (2011a,b)). Evaluation of the model simulations from GLDAS against in-situ and remote sensing measurements are presented in Rodell et al. (2004a) and Kato et al. (2007). The LandFlux-EVAL project, a more recent initiative, evaluated evapotranspiration estimates from a number of LSMs against in-situ data based estimates (Jiminez et al. (2011)). Approaches to define a minimum acceptable performance benchmark of LSMs by comparing them to calibrated noncausal (statistical/correlational) models are explored in Abramowitz et al. (2008). Though these efforts cover a wide spectrum of model evaluation and benchmarking of model process advancements, the evaluation criteria and the performance metrics tend to be specific to each application. LVT consolidates the requirements identified in these efforts within a single framework.

A number of software environments for conducting model verification has been reported in the literature. The Ensemble Verification System (EVS; Brown et al. (2010)) developed at the U.S. National Oceanic and Atmospheric Administration's (NOAA) Office of Hydrologic Development (OHD) provides an environment to verify ensemble forecasts of hydrologic and atmospheric variables such as precipitation, temperature and streamflow and is used by forecasters at the U.S. River Forecast Centers (RFCs). Protocol for the Analysis of Land Surface models (PALS) is a web-based application for evaluating land surface models against observed datasets and calibrated statistical models (Abramowitz et al. (2008)). LVT and PALS will continue to be developed concurrently to address community goals for benchmarking and MDF. Model Evaluation Toolkit (MET; Brown et al. (2009)) is a system developed by the Developmental Testbed Center (DTC) for the numerical weather prediction community to evaluate model performance. MET includes several methods for the diagnostic and spatial verification of NWP model outputs. However, MET requires that the input datasets (model output and the observational data) be reformatted to certain predefined file formats. LVT shares many features with these existing environments, but focuses on the native use of observational and model data sets since the interpretation of the data formats and reporting procedures is a critical and time consuming step in the evaluation process. LVT is designed as a framework that can be directly used and extended by the individual users and also includes a number of advanced features such as the evaluation of data assimilation diagnostics, standardized land surface diagnostics and uncertainty and information theory based analysis features. The following sections describe the design and capabilities of LVT.

3 Design of the LVT framework

LVT is implemented using object oriented framework design principles as a modular, extensible and reusable system. The software architecture of the system follows a three layer structure, as shown in Figure 2. LVT core, the top layer, encompasses generic modeling features such as the management of time, I/O, configuration, logging and geospatial transformations. The middle layer, called

"Abstractions" represents the extensible interfaces defined for incorporating additional functionalities into LVT. These include plugin interfaces for implementing new observational data sources and analysis metrics. The Abstractions layer provides the entry points for the reuse of existing generic capabilities of the LVT core. The top two layers thus represent the classic "semi-complete" nature of an object oriented framework, which is made fully functional by including specific implementations of the abstractions. As shown in Figure 2, implementations to read and process observations from a wide range of terrestrial hydrological observations have been implemented using the "*Observations*" abstraction. Similarly, a large suite of analysis metrics has been implemented by extending the "*Metrics*" abstraction.

LVT software is primarily written in Fortran 90 programming language. Though Fortran 90 lacks the direct support for object oriented programming concepts such as polymorphism and inheritance, these properties can be simulated in software (Decyk et al. (1997)) through the combined use of Fortran 90 and C programming languages. The compile-time polymorphism in LVT is simulated through the use of virtual function tables, by employing C language to interface with Fortran 90 functions, and by storing them in memory to be invoked at runtime.

A key advantage of this object oriented-based design is interoperability. The top two layers (LVT core and Abstractions) define the interactions between an *Observation* or a *Metric* implementation with the LVT core in a generic manner. Similarly, the required interconnections between an *Observation* implementation and a *Metric* implementation are also handled generically. As a result, the existing functionalities of the system are automatically available to a new addition in LVT, implemented through the extension of an Abstraction. For example, a newly incorporated observation implementation can take advantage of all available analysis metrics without having to define any additional interconnections between each bottom layer component.

Note that many of the model-independent capabilities within the LVT are enabled by the Earth System Modeling Framework (ESMF; Hill et al. (2004)). ESMF provides a structured collection of building blocks that can be customized to develop model components for Earth Science applications. It provides an infrastructure of utilities and a superstructure for coupling different model components. LVT employs the ESMF infrastructure utilities to handle the management of clock/time, configuration, and logging. Further, LVT also employs the generic ESMF objects (called ESMF States) for sharing data and information between different components.

4 Capabilities of LVT

A critical part of an evaluation procedure is the processing of datasets, which normally consists of model outputs and measurements from in-situ, satellite and remote sensing platforms. These datasets typically have different file formats, spatial and temporal scales and reporting procedures. Further, the in-situ and remotely sensed measurements typically require extensive quality control before their

use. The rectification of such differences between datasets being compared is an essential, but routine and time consuming step in the evaluation process. The philosophy in LVT is to use the datasets in their native formats. The "plugin" style design of LVT enables the development of data processors corresponding to each dataset. Once developed, these data processors can be subsequently used to work with an ongoing data collection without additional reprocessing.

4.1 Support for terrestrial hydrological datasets in LVT

The key processes that constitute the terrestrial hydrological cycle include precipitation, radiation, interception of precipitation by vegetation, infiltration of precipitation into the soil and the vertical transfer of soil moisture, evapotranspiration, formation of snow, snow melt, and river runoffs, among others. In order to quantify the contribution of these individual processes to the overall variability of the terrestrial hydrological cycle, they must be evaluated against the full suite of available measurements. Motivated by this goal, the processing of a large set of measurements of different processes from a variety of sources are supported in LVT. As shown in Table 1, these datasets constitute the monitoring of different components of the terrestrial hydrological cycle, from different observing platforms. The spatial and temporal scales of these measurements also vary significantly. By incorporating the processing of these datasets under a single, integrated framework, LVT enables an environment for performing a comprehensive evaluation of the terrestrial hydrological processes. Note that the support of this large suite of products is enabled by the extensible nature of LVT software design and is expected to further expedite the incorporation of other relevant datasets in the future.

4.2 Analysis Metrics

The need for having a variety of performance evaluation metrics in the verification process is well recognized (Stanski et al. (1989)), as the robustness and sensitivity of each metric to measurement attribute vary (Entekhabi et al. (2010)). Further, the appropriateness of an analysis metric may also differ significantly based on the targeted application (Gupta et al. (2009)). Model evaluation studies quite often use accuracy-based metrics that quantify model performance using residual-based measures. These metrics, however, may not provide further insights on the robustness of the model under future or unobserved scenarios (Pachepsky et al. (2006)). They are also inadequate in capturing estimates of associated uncertainties (Gulden et al. (2008)), relative importance and sensitivity of model parameters to the overall accuracy and uncertainty, tradeoffs in performance due to spatial scales and the tradeoffs between actual information content and variabilities introduced by random noise. Gupta et al. (2008) emphasize the need for sophisticated diagnostic evaluation methods that help in isolating the limitations of the model representations.

A number of analysis metric types is supported in LVT including; (1) Statistical accuracy measures that are conventionally used for model evaluation by comparing the model simulation against

independent measurements and observations (e.g. RMSE, Bias), (2) Ensemble measures that provide assessments of the accuracy of probabilistic model outputs against observations, (3) Metrics that help in quantifying the apportionment of uncertainty and sensitivity of model simulations to model parameters, (4) Information theory-based measures that provide estimates of information content and complexity associated with model simulations and measurements, (5) Spatial similarity and scale decomposition methods that assist in quantifying the impact of spatial scales in model improvements and errors and (6) Standard diagnostics to evaluate the efficiency of computational algorithms such as data assimilation. Table 2 presents a list of supported metric implementations within LVT. The details of the metric implementations are discussed in Section 5 through a number of illustrative examples. The availability of this suite of metrics enables novel ways to quantify and translate model performance.

4.3 Miscellaneous features

LVT also supports a number of miscellaneous features to assist the verification procedures. To provide a measure of the statistical significance and the influence of sampling density on the results, confidence intervals based on Gaussian distributions are computed for each verification metric. LVT generates the results of the analyses in ASCII text, binary, GriB and NetCDF output formats. The capabilities to generate probability density functions (PDFs) of the computed metrics by stratifying to specified parameters are also included in LVT. Further, LVT also provides methods to impose user-defined masking to exclude selected grid points when analysis metrics are computed. These masks can be static, time-varying or based on a certain variable. For e.g., a downward shortwave radiation ($SW\downarrow$) based mask can be defined that separates the analysis computations when the $SW\downarrow$ values are above and below a specified threshold (say 5 W/m^2). This will enable a day-night stratification of the computed metrics, when $SW\downarrow$ values are above and below 5 W/m^2, respectively.

LVT also includes a number of land surface process diagnostics related to the partitioning of energy across the land atmosphere interface such as evaporative fraction, bowen ratio and overall energy, water and evaporation budgets at the land-atmosphere interface. These diagnostics are computed for both model and observational datasets. Quantifying these diagnostics are important for improving the understanding of the feedbacks between the land surface and the atmosphere.

As mentioned earlier, LVT also supports the analysis of diagnostics generated by the LIS data assimilation subsystem. These include distribution statistics of data assimilation innovations and analysis gain, which provide measures of the efficiency of data assimilation configurations. Similarly, LVT also handles the outputs of the optimization and uncertainty estimation subsystems of LIS. For e.g., checks to assess the convergence of these iterative algorithms can be performed by analyzing the optimization and uncertainty estimation outputs through LVT.

Though LVT was originally designed to support LIS outputs, it has since been extended to facilitate the evaluation of other "non-LIS" model products. LVT contains the features to convert the

given non-LIS product to a LIS output style and format. It then uses the converted output for evaluation. Note that this process does not involve any spatial or temporal transformation of the data, rather the conversion to a different data format and convention.

5 Model evaluation examples using LVT

5.1 An end-to-end example of the MDF paradigm

As noted earlier, one of the key motivations behind LVT is to provide a system that can augment LIS' modeling capabilities with an evaluation framework. The joint use of both these systems enables an end-to-end environment for facilitating the steps of the MDF paradigm. In this section, we present an example of using the modeling and computational tools in LIS to refine the model performance and the verification features in LVT to quantitatively evaluate the simulations.

Model simulations using the Noah LSM (version 3.2) (Ek et al. (2003); Barlage et al. (2010)) forced with the NLDAS-II datasets are conducted over a 500x500 domain covering the U.S. Southern Great Plains (SGP) at 1km spatial resolution during the time period of 1 May, 2006 to 1 September, 2006. This domain is used in a number of prior studies on land-atmosphere feedbacks (Santanello et al. (2009, 2011)). Using the default values of the soil and vegetation parameters of the Noah LSM, a model simulation is conducted first to simulate surface latent and sensible heat flux estimates. Using LVT, these flux estimates are evaluated against the in-situ measurements from 19 Atmospheric Radiation Measurement (ARM) stations. The optimization algorithms in LIS are then used to estimate a refined set of model parameters with the objective of minimizing the cumulative error in the hourly surface flux observations from the ARM stations, over the four month period. Subsequently, the improved model performance with the calibrated parameters is quantified using LVT.

Figure 3 shows a comparison of the mean diurnal cycles of latent and sensible heat fluxes from model simulations compared against that of the measurements from 19 ARM-SGP stations. The simulations using default model parameters show large errors, with a significant underestimation in the latent heat fluxes and an overestimation in sensible heat fluxes. The calibration of model parameters helps in improving the model performance, by correcting both these systematic biases. This example illustrates an example of the MDF paradigm that includes model characterization, reformulation through parameter estimation, and verification using LVT. Similar instances can be implemented using the extensive evaluation capabilities of LVT.

5.2 Example of model evaluation against satellite data

Model formulation and evaluation are typically conducted over instrumented locations of the world where independent measurements are available. Though these in-situ observations provide valuable information on the spatial and temporal variability of process variables, they are limited in their spatial coverage. Satellite and remotely-sensed measurements, on the other hand, have improved

spatial coverages and they enable the extension of model evaluation to uninstrumented locations and hydrologic regimes. In this section, we present an example of model evaluation against satellite data over a region where in-situ measurements are sparse.

A model simulation using Noah LSM (version 2.7.1) is conducted over a 1200km x 1000km domain, at 1km spatial resolution over Afghanistan from 1 Oct 2007 to 1 May 2010. The LSM is driven with meteorological data from the Global Data Assimilation System (GDAS); the global meteorological weather forecast model of the National Centers for Environmental Prediction (Derber et al. (1991)). The precipitation input for the model simulations is provided from the NOAA Climate Prediction Center's (CPC) operational global 2.5° 5-day Merged Analysis of Precipitation (CMAP; Xie and Arkin (1997)), which is a product that employs blended satellite (IR and microwave) and gauge observations. The model domain has complex terrain characteristics, with elevation ranges from 1000 to 6000 m. The fractional snow cover extent global 500m product (MOD10A1 Version 4; Hall et al. (2006)) from the Moderate Resolution Imaging Spectroradiometer (MODIS) optical sensor on the Terra spacecraft is used as the reference data for evaluating simulations of snow cover fields simulated by the LSM. The MOD10A1 product is aggregated to 1km spatial resolution for enabling the comparisons presented here.

The snow cover fields are evaluated by computing the probability of detection (POD) and false alarm ratio (FAR) against the MOD10A1 product. POD measures the fraction of snow cover presence that were correctly simulated and FAR quantifies the fraction of no-snow events that were incorrectly simulated. Figure 4 shows the average POD and FAR values during the model simulation period, computed using detection threshold of 0.8 (above which a positive detection of snow cover simulation is assumed). The POD and FAR fields display the terrain features of the Hindu Kush mountains, that run northeast to southwest. High values of POD and low values of FAR are observed over the Central Highlands region of the domain, suggesting a high degree of accuracy of model snow cover estimates over these areas. Over the northeast parts of the domain, however, the model simulations are less accurate, as indicated by the lower POD and higher FAR values.

5.3 Analysis of data assimilation diagnostics

The example in Section 5.1 presents an instance of the MDF paradigm that employs parameter estimation for model reformulation. As noted in Williams et al. (2009), similar MDF instances can be defined that employ data assimilation techniques to improve state estimation. This section presents an example of using data assimilation diagnostics to assess the performance of the system within a MDF context.

The difference between the observations being assimilated and the model forecasts, known as innovations, are typically computed during data assimilation. The statistics of the innovations are typically used to diagnose the performance of the assimilation algorithm. For example, when the Ensemble Kalman Filter (EnKF) is used as the assimilation algorithm, a linear system dynamics is

assumed with Gaussian, mutually and serially uncorrelated errors in model and observations (Reichle and Koster (2002)). Consequently, the distribution of normalized innovations (normalized with their expected covariance) is expected to follow a standard normal distribution $N(0,1)$ (Gelb (1974)). The deviations from the expected mean and standard deviation of the normalized innovation distribution is used as a measure of suboptimality of the data assimilation configuration. A number of studies have confirmed that poor specification of model and observation error parameters can significantly degrade the quality of assimilation products (Reichle and Crow (2008); Reichle et al. (2008)). The assimilation diagnostics can be analyzed using LVT and the model and observation error specifications can then be continually revised to ensure optimal data assimilation performance.

To demonstrate these capabilities, a synthetic data assimilation experiment is conducted over the Continental U.S. domain at 1° spatial resolution, for a time period of 1 Jan 2000 to 1 Jan 2006. In this experiment, the observations to be assimilated are synthetically simulated (from an independent land model simulation using the Catchment LSM) and as a result, the associated errors are perfectly known. The observations are assimilated using the Ensemble Kalman Filter (EnKF) algorithm. The details of the assimilation setup is provided in Kumar et al. (2011). Figure 5 shows the spatial distribution of mean and variance of normalized innovations over the domain generated by the assimilation system. In this instance, the mean values are close to zero and the variances are closer to 1, indicating the near-optimal performance. Additional analysis metrics such as lag correlation coefficients to assess the "whiteness" of the innovation distribution are also provided within LVT for more detailed evaluations of the efficiency of the data assimilation system.

5.4 Characterization of uncertainty diagnostics

It is well acknowledged that model simulations and observations are affected by different sources of uncertainties. The errors in model parameters, input forcing and structural deficiencies introduce uncertainties in the model simulations. The measurements from satellite and remote sensing platforms are subject to measurement noise and errors in retrieval models. Similarly, the in-situ measurements also have associated uncertainties due to environmental factors, data processing and instrument errors. Therefore, it is important to quantify the impact of these uncertainty sources in modeled estimates. LVT includes a number of measures to quantify the propagation of model parameter uncertainty in predictions.

To demonstrate the use of uncertainty analysis metrics, a model simulation using Noah LSM (version 3.2) is conducted during the summer months (May to September) of 2010 over a region encompassing the Walnut Gulch watershed in southeastern Arizona. The meteorological boundary conditions from the Agricultural Meteorology Model (AGRMET; Moore et al. (1990)) are used to force the models at 0.25° spatial resolutions. The in-situ measurements of soil moisture values are used to evaluate the model simulations. To investigate the impact of parameter uncertainty in simulated soil moisture estimates, a Monte Carlo (MC) simulation is conducted by sampling four soil

hydraulic properties (SHPs) (θ_s - porosity, ψ_s - saturated matric potential, K_s - saturated hydraulic conductivity and b - pore size distribution index) from assumed uniform distributions. The simulation uses an ensemble size of 100. Figure 6(a) shows a time series comparison of the model simulation of surface soil moisture against the in-situ measurements. Note that the vertical profile of observations are suitably weighted to provide an equivalent comparison against the model simulation which represents a surface layer of 10 cm depth. The comparison indicates significant differences between the ensemble mean and the observations. Further, the consideration of uncertainty in SHPs translates to significant uncertainty in simulated soil moisture. The shaded region (shown as $\pm$ 2$\times$ the ensemble standard deviation) around the ensemble mean represents the uncertainty in simulated soil moisture. The soil moisture uncertainty is small during the dry period, but grows significantly during the late summer months when both the magnitude and variability of soil moisture increase. Though the spread of the ensemble encompasses the observations, the observations tend to fall towards the tail end of the ensemble distribution. This emphasizes the need to refine the model parameters and their sampling strategies for a better characterization of modeling uncertainty.

Figure 6(b) also provides an uncertainty importance measure which is an assessment of the relative contribution of each parameter to the ensemble spread. This metric is computed as the correlation between the simulated variable (surface soil moisture) and the parameter across the ensemble. Figure 6(b) suggests that among the four SHPs considered, model simulations are most sensitive to θ_s, followed by K_s. The variability in ψ_s and the b parameters contribute less to the uncertainty in soil moisture in this instance. The figure also illustrates that the relative importance of the parameter is sensitive to the soil moisture magnitude and variability. During the late summer months, the uncertainty importance of θ_s also increases with the magnitude of simulated soil moisture. Knowledge of the relative importance of the model parameters is significant when choosing the set of model parameters for calibration and sampling, and LVT facilitates the quantification such sensitivities. Similar to the examples described in Sections 5.1 and 5.3, this example provides another instance of using LVT to enable the MDF concept, in the context of uncertainty estimation.

5.5 Information Theory metrics

A number of studies (Wackerbauer et al. (1994); Lange (1999); Selle and Huwe (2004)) describe the use of information theory-based metrics to discriminate time series data based on their information content (or randomness) and their complexity. Pachepsky et al. (2006) and Pan et al. (2011) describe the use of these measures for discriminating soil water models. LVT includes a number of information theory-based measures such as metric entropy, mean information gain, effective complexity and fluctuation complexity. These measures are computed by converting the time series of a given dataset into a binary symbol string (Lange (1999)). Within the symbol string, patterns of words (defined as a group of consecutive symbols of a certain length) are identified, representing a state of the system of interest. For e.g., a word consisting of L consecutive symbols has 2^L possible states.

The information theory metrics are then defined by computing the probabilities associated with the patterns of words in the converted time series of the data. For example, the metric entropy (ME) and information gain (IG) metrics are defined as follows:

$$ME = -\frac{1}{L}\sum_{i=1}^{2^L} p_i log_2 p_i \tag{1}$$

$$IG = -\sum_{i,j=1}^{2^L} p_{L,ij} log_2 p_{L,i\rightarrow j} \tag{2}$$

where p_i is the probability of occurrence of the ith word, $p_{L,ij}$ is the probability of transition from the ith to the jth word, and $p_{L,i\rightarrow j}$ is the conditional probability of the occurrence of the jth word given that the ith word has already occurred in the symbol sequence. A more detailed description of these measures are provided in Pachepsky et al. (2006).

The information theory-based metrics are typically applied to discriminate model simulations, especially when they yield similar accuracy measures. Here we demonstrate their use for comparing soil moisture simulations from Noah LSM (version 3.2) when two different retrievals from the Advanced Microwave Scanning Radiometer for the Earth Observing System (AMSR-E) sensor aboard the Aqua satellite are assimilated. The NASA Level-3, "AE_Land3" product (version 6,Njoku et al. (2003)) and the AMSR-E Land Parameter Retrieval Model (LPRM) product developed at NASA GSFC and VU Amsterdam (Owe et al. (2008)) are used in the data assimilation (DA) integrations. The experiments are carried out over the Continental United States for a period of 2002 to 2008, using the same configuration used in the NLDAS project (Mitchell et al. (2004)) (from 25-53°N and 125-67°W at 1/8 degree spatial resolution). The details of the assimilation methodology are described in Peters-Lidard et al. (2011).

Figure 7 presents a comparison of the change in metric entropy (ΔME) and the information gain (ΔIG) metric as a result of data assimilation. These metric values are computed using a word length of 3. The ΔME and ΔIG values are calculated by subtracting the metric values for the simulation without data assimilation from the corresponding data assimilation integration. Figure 7 indicates that DA introduces more entropy (randomness) in the simulations, over most parts of the domain, with higher values of ΔME for the NASA DA compared to the LPRM DA. The information gain metric indicates how much the sequence of patterns in the data contributes to the overall information. The ΔIG values when assimilating NASA retrievals are larger compared to that of LPRM assimilation. The changes in soil moisture introduced by the NASA DA also result in more randomness in the consecutive patterns in the time series. This leads to higher IG values for NASA DA relative to LPRM DA, suggesting that the changes in soil moisture time series introduced by LPRM DA may be less spurious (random). In prior MDF studies (Reichle et al. (2007); Liu et al. (2011a); Peters-Lidard et al. (2011)) accuracy-based measures were used to characterize the value of assimilating these retrievals in to LSMs. The results in this article present an alternate evaluation

using information theory metrics within LVT.

5.6 Scale decomposition features

Study of the effects of spatial scale has been an active area of hydrological research (Gupta et al. (1986); Wood et al. (1990); Sivapalan and Kalma (1995); Seyfried and Wilcox (1995); Bloschl and Sivapalan (1995); Wood et al. (1988); Bloschl (1999); Erickson et al. (2005); Trujillo et al. (2009)). Characterization of the nature of spatial variability of different component processes over a range of scales are important for improving the utility of terrestrial hydrological models. LVT includes approaches such as discrete wavelet transforms to enable scale based decomposition analyses. Here we present an example of scale-decomposition evaluation of snow cover simulations from the LSMs using LVT.

The intensity-scale approach of Casati et al. (2004), originally developed for the spatial verification of precipitation forecasts, is used to perform a scale decomposition analysis. The technique employs a two dimensional discrete Haar wavelet transform that decomposes a given field into sum of orthogonal components at different spatial scales. The mean squared error (MSE) of the decomposed components at each spatial scale is used to quantify the scale decomposition effects.

Using the domain configuration at 1km spatial resolution over Afghanistan used in Section 5.1, two model simulations are conducted using Noah LSM (version 2.7.1); one that employs a terrain based correction of shortwave radiation input to the LSM and one that does not include such adjustments. The terrain-based corrections adjust the incoming shortwave radiation based on terrain slope and aspect and these changes in turn impact the evolution of snow over these terrain. The improvements in the snow cover simulation as a result of the terrain-based correction is computed as the difference in POD fields from the two simulations, generated by comparing against the MOD10A1 (version 4) fractional snow cover product. The scale-decomposition approach is then applied to this difference field to quantify how the improvements in snow cover estimates at 1km spatial resolution translate to coarser spatial scales.

Figure 8 shows the result of scale decomposition of the total improvement field for POD using the two dimensional discrete Haar wavelet transform. The algorithm computes successive decompositions of the original field by powers of 2. The percentage contribution to the total improvement at each coarse spatial scale is shown in Figure 8. The results indicate that most of the improvements in POD are obtained at fine spatial scales and the contribution of the scale decreases with increase in spatial resolution. At scales coarser than 16km, the percentage contribution drops below 10%. Similar analysis of scale effects can be performed on other metrics and variables of interest. This example demonstrates the use of LVT for another MDF experiment where the MODIS fractional snow cover data is used to assess the applicability of model formulations at different spatial scales.

5.7 Spatial similarity measures

With the increased availability of spatially distributed datasets from satellites and remote-sensing platforms, there is a need for techniques and metrics that evaluate models and observations based on the their spatial patterns, in addition to the one-to-one correspondence comparisons that are typically used. The incorporation of spatial pattern comparisons will aid in further improving the reliability of LSMs for hydrological applications (Bloschl and Sivapalan (1995); Grayson and Bloschl (2000)). A review of spatial similarity methods in hydrology is provided in Wealands et al. (2005), which includes techniques based on statistical identification as well as image processing techniques. In this section, an example of using a similarity metric through LVT to compare snow cover patterns from two different LSMs is presented.

Snow cover estimates using two LSMs, Noah (version 3.2) and CLM (version 2 ; Dai et al. (2003)), forced with GDAS and CMAP datasets, are generated over a 100x100 region near the Southern Great Plains in the US at 1km spatial resolution for a time period of November 1, 2008 to 1 June 2009. The LSMs have different representation of snow processes, with Noah employing a simple single snow layer scheme. CLM includes a more complex five layer snow scheme with parameterizations for temporally varying snow albedo, as a function of snow cover and snow age. Both LSMs simulate temporally varying snow density with evolution of patchy snow cover. The model simulations are evaluated against the fractional snow cover observations from MODIS (MOD10A1 version 4) using the "Hausdorff distance" similarity metric.

Hausdorff distance (HD) measures the similarity of points in two finite sets and is not designed to find one-to-one correspondence between points in each set. It is expressed as the maximum distance of a set to the nearest point in the other set.

$$h(M,O) = \max_{m \in M} \{ \min_{o \in O} \{ ||m - o|| \} \} \tag{3}$$

where $h(M,O)$ is the HD value, m and o are points of sets M (representing model) and O (representing observations), respectively. $||m-o||$ is the norm of the points in the model and observation spaces and can be computed as the Euclidean distance between m and o.

Figure 9 shows a time series comparison of the cumulative HD measure from Noah and CLM snow cover simulations for the winter season of 1 November, 2008 to 1 June, 2009. More temporal variability in HD values is observed during the snow evolution and ablation periods and it drops during the peak snow season, suggested by the flattening of the cumulative HD curves. This indicates that there is more consistent agreement in the observational and model simulated patterns during the peak snow season. During the snow melt period, Noah produces lower HD values compared to CLM. This suggests that the spatial patterns in the Noah snow cover simulations capture the observational patterns more accurately relative to CLM's simulations, though CLM's snow physics formulations are more complex. Note that newer versions of both these models (Noah-MP (Niu et al. (2011)) and CLM version 4.0 (Lawrence et al. (2011))) with updated snow physics formulations are currently

being incorporatd into LIS and similar comparisons can be performed through LVT to evaluate the updated snow physics in these LSMs. This experiment demonstrates the use of spatial similarity metrics for comparing the performance of two different LSMs within a MDF framework.

6 Summary and Future Directions

This article describes the development and capabilities of a verification system for terrestrial hydrology known as the Land surface Verification Toolkit. LVT enables an environment for conducting the systematic evaluation of land model outputs by providing a variety of analysis metrics and procedures. LVT functions primarily as an analysis back-end system for the NASA Land Information System (LIS), but also supports the analysis of data products from other modeling environments. LIS is a comprehensive land surface modeling framework and includes data assimilation and posterior inference tools such as optimization and uncertainty estimation to facilitate the exploitation of information content from observational datasets to augment model predictions. LVT not only supports the verification of LSM outputs, but also provides the tools to analyze the performance of these computational algorithms within LIS. LVT is designed using object oriented software principles, with abstractions defined for the customization and extension of the system for different applications. These extensible interfaces allow the incorporation of new observational datasets and analysis metrics in an interoperable manner. The combination of the modeling capabilities of LIS and the analysis capabilities of LVT provide a robust environment for conducting end-to-end model data fusion experiments that has been identified in the community as a key paradigm for improving the applicability of LSMs.

LVT currently supports a large suite of in-situ, satellite and remotely-sensed, and model and reanalysis products to enable comprehensive evaluations of various hydrological processes. These datasets are supported in their native format and LVT handles the temporal and spatial transformations required in the analysis. Diagnostic model verification and intercomparisons are supported through a variety of analysis metrics and procedures. In addition to the standard accuracy-based measures, LVT supports ensemble and uncertainty measures, metrics based on information theory, similarity metrics and methods to quantify the impact of spatial scales on model performance. This variety of techniques provide novel ways to characterize model performance and to investigate associated tradeoffs.

The article presents a number of illustrative examples that demonstrate the capabilities of LVT and provide several instances of end-to-end MDF experiments. The optimization algorithms in LIS are used to refine the model parameters of the LSM to improve its estimation of surface fluxes. LVT is used to quantify the systematic improvements resulting from the refined model parameters. The impact of data fusion for model state and uncertainty estimation is assessed through data assimilation and uncertainty quantification metrics, respectively. The information theory-based metrics

provide measures such as metric entropy, information gain and complexity to identify tradeoffs in datasets based on their information content and complexity. Acknowledging the need to perform model evaluations in a spatially distributed manner, spatial similarity metrics and scale decomposition techniques that provide spatial pattern comparisons against remotely-sensed distributed datasets are also incorporated in LVT.

LVT is an evolving framework and continues to be enhanced with the addition of new analysis capabilities and the incorporation of terrestrial hydrological datasets. In addition to the handling of LSM outputs, the support for outputs from various application models coupled to LIS (e.g. crop, drought, flood, landslide models) is also being developed. Ensemble measures such as reliability, resolution and discrimination (Murphy and Winkler (1992)) and timing error measures (Liu et al. (2011b)) will also be incorporated into the current suite of analysis metrics. The use of a common environment for diagnostic evaluation will also help in quantifying the tradeoffs between different metrics and skill scores. For e.g., different organizations use different indices for quantifying the severity of drought (Heim (2002)). The availability of these drought indices through LVT will enable cross-comparisons of these measures and the assessment of their suitability for the intended application. In summary, the growing capabilities of LVT are expected to help in the definition and refinement of a formal benchmarking and evaluation process for the LSMs and assist in improving their use for real-world applications.

Acknowledgements. We gratefully acknowledge the financial support from NASA Earth Science Technology Office (ESTO) and the US Air Force Weather Agency (AFWA) and the assistance from summer intern students Teodor Georgiev (Princeton University), Yi Yuan (University of Michigan) and Corina Robles (Florida International University) for assembling evaluation datasets and the software testing of LVT. Computing was supported by the resources at the NASA Center for Climate Simulation.

References

Abramowitz, G., Leuning, R., Clark, M., and Pitman, A.: Evaluating the performance of land surface models, J. Climate, 21, 5468–5481, 2008.

Barlage, M., Chen, F., Tewari, M., Ikeda, K., Gochis, D., Dudhia, J., Rasmussen, R., Livneh, B., Ek, M., and Mitchell, M.: Noah Land Surface Model modifications to improve snowpack prediction in the Colorado Rocky Mountains, Journal of Geophysical Research, 115, doi:10.1029/2009JD013470, 2010.

Barrett, A.: National Operational Hydrologic Remote Sensing Center Snow Data Assimilation System (SNODAS) products at NSIDC, Tech. rep., National Snow and Ice Data Center, Boulder, CO, digital Media, 2003.

Bloschl, G.: Scaling issues in snow hydrology, Hydrological Processes, 13, 2149–2175, 1999.

Bloschl, G. and Sivapalan, M.: Scale issues in hydrological modeling, Hydrological Processes, pp. 251–290, 1995.

Blyth, E., Gash, J., Lloyd, A., Pryor, M., Weedon, G., and Shuttleworth, J.: Evaluating the JULES model energy fluxes using the FLUXNET data, J. Hyrometeor., 11, 509–519, 2010.

Blyth, E., Clark, D., Ellis, R., Huntingford, C., Los, S., Pryor, M., Best, M., and Sitch, S.: A comprehensive set of benchmark tests for a land surface model of simultaneous fluxes of water and carbon at both the global and seasonal scale, Geosci. Model Dev., 4, 255–269, doi:10.5194/gmd-4-255-2011, 2011.

Brown, B., Gotway, J., Bullock, R., Gilleland, E., Fowler, T., Ahijevych, D., and Jensen, T.: The Model Evaluation Tools (MET): Community tools for forecast evaluation, in: 25th Conf. on International Interactive Information and Processing Systems (IIPS) for Meteorology, Oceanography, and Hydrology, Amer. Metero. Soc., Phoenix, AZ, 2009.

Brown, J., Demargne, J., Seo, D.-J., and Liu, Y.: The Ensemble Verification System (EVS): a software tool for verifying ensemble foreasts of hydrometeorological and hydrologic variables at discrete locations, Environmental Modeling and Software, 25, 854–872, 2010.

Brown, R. and Brasnett, B.: Canadian Meteorological Center (CMC) daily snow analysis data, Tech. rep., National Snow and Ice Data Center, Boulder, CO, digital Media, 2010.

Casati, B., Ross, G., and Stephenson, D.: A new intensity-scale approach for the verification of spatial precipitation forecasts, Meteorol. Appl., 11, doi:10.1017/S1350482704001239, 2004.

Chen, F., Mitchell, K., Schaake, J., Xue, Y., Pan, H., Koren, V., Duan, Y., Ek, M., and Betts, A.: Modeling of land-surface evaporation by four schemes and comparison with FIFE observations, J. Geophys. Res., 101, 7251–7268, 1996.

Dai, Y., Zeng, X., Dickinson, R., Baker, I., Bonan, G., Bosilovich, M., Denning, S., Dirmeyer, P., Houser, P., Niu, G., Oleson, K., Schlosser, A., and Yang, Z.-L.: The common land model (CLM), Bulletin of the American Meteorological Society, 84, 1013–1023, dOI:10.1175/BAMS-84-8-1013, 2003.

de Rosnay, P., Boone, A., Beljaars, A., and Polcher, J.: AMMA Land surface intercomparison projects, GEWEX News, 16, 10–11, 2006.

Decyk, V. K., Norton, C. D., and Szymanski, B. K.: How to express C++ concepts in Fortran 90, Scientific Programming, 6, 363–390, 1997.

Derber, J., Parrish, D., and Lord, S.: The new global operational analysis system at the National Meteorological Center, Weather and Forecasting, 6, 538–547, 1991.

Dirmeyer, P., Gao, X., Zhao, M., Guo, Z., Oki, T., and Hanasaki, N.: GSWP-2: Multimodel analysis and implications for our perception of the land surface, Bull. Amer. Meteor. Soc., 87, 1381–1397, doi:10.1175/BAMS-87-10-1381, 2006.

Ek, M., Mitchell, K., Yin, L., Rogers, P., Grunmann, P., Koren, V., Gayno, G., and Tarpley, J.: Implementation of Noah land-surface model advances in the NCEP operational mesoscale Eta model, Journal of Geophysical Research, 108, doi:10.1029/2002JD003 296, 2003.

Entekhabi, D., Asrar, G., Betts, A., Beven, K., Bras, R., Duffy, C., Dunne, T., Koster, R., Lettenmaier, D., McLaughlin, D., Shuttleworth, W., van Genuchten, M., Wei, M.-Y., and Wood, E.: An agenda for land surface hydrology research and a call for the second international hydrological decade, Bull. Amer. Meteor. Soc., 80, 2043–2058, 1999.

Entekhabi, D., Reichle, R., Koster, R., and Crow, W.: Performance Metrics for Soil Moisture Retrievals and Application Requirements, J. Hydrometeor., 11, 832–840, doi:doi:10.1175/2010JHM1223.1, 2010.

Erickson, T., Williams, M., and Winstral, A.: Persistence of topographic controls on the spatial distribution of snow in rugged mountain terrain, Colorado, United States, Water Resources Research, 41, 10.1029/2003WR002973, 2005.

Fennessey, M. and Shukla, J.: Impact of initial soil wetness on seasonal atmospheric prediction, J. Climate, 12, 3167–3180, 1999.

Foster, J., Hall, D., Eylander, J., Riggs, G., Nghiem, S., M.Tedesco, E.Kim, Montesano, P., Kelly, R., K.A.Casey, and B.Choudhury: A blended global snow product using visible, passive microwave and scatterometer satellite data, International Journal of Remote Sensing, 32, doi:10.1080/01431160903548013, 2011.

Friend, A. and Kiang, N.: Land surface model development for the GISS GCM: Effects of improved canopy physiology on simulated climate, J. Climate, 18, 2883–2902, 2005.

Gelb, A.: Applied Optimal Estimation, MIT Press, Cambridge, MA, 1974.

Grayson, R. and Bloschl, G.: Spatial patterns in catchment hydrology: Observations and modeling, Cambridge University Press, Cambridge, 2000.

Gulden, L., Rosero, E., Yang, Z.-L., Wagener, T., and Niu, G.-Y.: Model performance, model robustness, and model fitness scores: A new method for identifying good land surface models, Geophys. Res. Lett., 35, 10.1029/2008GL033721, 2008.

Gupta, H., Wagener, T., and Liu, Y.: Reconciling theory with observations: elements of a diagnostic approach to model evaluation, Hydrological Processes, 22, 3802–3813, doi:10.1002/hyp.6989, 2008.

Gupta, H., Kling, H., Yilmaz, K., and Martinez, G.: Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modeling, J. Hydrology, 377, 80–91, doi:10.1016/j.jhydrol.2009.08.003, 2009.

Gupta, V., Rodriguez-Iturbe, I., and Wood, E.: Scale problems in Hydrology, Reidel, Dordrecht, 1986.

Hall, D., Riggs, G., and Salomonson, V.: MODIS/Terra snow cover daily L3 Global 500m Grid V005, Tech. rep., National Snow and Ice Data Center, Colorado, USA, digital media, 2006.

Harrison, K., Kumar, S., Peters-Lidard, C., and Santanello, J.: Quantifying soil moisture modeling uncertainty with remote sensing observations using Bayesian inference techniques, Water Resources Research, in preparation, 2011.

Heim, R. J.: A review of twentieth century drought indices used in the United States, Bull. Amer. Meteor. Soc., 83, 1149–1165, 2002.

Henderson-Sellers, A., Pitman, A., P.K., L., Irannejad, P., and Chen, T.: The project for Intercomparison of land surface parameterization schemes (PILPS): Phases 2 and 3, Bull. Amer. Meteor. Soc., 76, 489–503, 1995.

Higgins, R., Janowiak, J., and Yao, Y.-P.: A gridded hourly precipitation database for the United States (1963-1993)., Tech. rep., NCEP Climate Prediction Center Atlas 1, 46pp, 1996.

Hill, C., DeLuca, C., Balaji, V., Suarez, M., and da Silva, A.: The Architecture of the Earth System Modeling Framework, Computing in Science and Engineering, 6, 2004.

Jiminez, C., Prigent, C., Mueller, B., Seneviratne, S., McCabe, M., Wood, E., Rossow, W., Balsamo, G., Betts, A., Dirmeyer, P., Fisher, J., Jung, M., Kanamitsu, M., Reichle, R., Reichstein, M., Rodell, M., Sheffield, J., Tu, K., and Wang, K.: Global intercomparison of 12 land surface heat flux estimates, Journal of Geophysical Research, 116, 2011.

Jung, M., Reichstein, M., and Bondeau, A.: Towards global empirical upscaling of FLUXNET eddy covariance observations: validation of a model tree ensemble approach using a biosphere model, Biogeosciences, 116, doi:10.1029/2010JD014545, 2009.

Kato, H., Rodell, M., Beyrich, F., Cleugh, H., van Gorsel, E., Liu, H., and Meyers, T.: Sensitivity of land surface simulations to model physics, land characteristics, and forcings at four CEOP sites, J. Meteor. Soc. Japan, 85A, 187–204, 2007.

Koren, V., Schaake, J., Mitchell, K., Duan, Q.-Y., and Chen, F.: A parameterization of snowpack and frozen ground intended for NCEP weather and climate models, J. Geophys. Res., 104, 19 569–19 585, 1999.

Koster, R., Suarez, M., Liu, P., Jambor, U., Berg, A., Kistler, M., Reichle, R., Rodell, M., and Famiglietti, J.: Realistic initialization of land surface states: Impacts on subseasonal forecast skill, J. Hydrometeor., 5, 1049–1063, 2004.

Koster, R., Guo, Z., Dirmeyer, P., Yang, R., Mitchell, K., and Puma, M.: On the nature of soil moisture in land surface models, J. Climate, 22, 4322–4335, doi:10.1175/2009JCLI2832.1, 2009.

Kumar, S., Peters-Lidard, C., Tian, T., Houser, P., Geiger, J., Olden, S., Lighty, L., Eastman, J., Doty, B., Dirmeyer, P., Adams, J., Mitchell, K., Wood, E., and Sheffield, J.: Land information system: An interoperable framework for high resolution land surface modeling, Environmental Modeling and Software, 21, 1402–1415, 2006.

Kumar, S., Peters-Lidard, C., Eastman, J. L., and Tao, W.-K.: An integrated high resolution hydrometeorological modeling testbed using LIS and WRF, Environmental Modelling and Software, 23, 169–181, 2007.

Kumar, S., Peters-Lidard, C., Tian, Y., Reichle, R. H., Alonge, C., Geiger, J., Eylander, J., and Houser, P.: An integrated hydrologic modeling and data assimilation framework enabled by the Land Information System (LIS), IEEE Computer, 41, 52–59, doi:10.1109/MC.2008.511, 2008a.

Kumar, S., Reichle, R., Peters-Lidard, C., Koster, R., Zhan, X., Crow, W., Eylander, J., and Houser, P.: A land surface data assimilation framework using the Land Information System: Description and Applications, Advances in Water Resources, 31, 1419–1432, doi:10.1016/j.advwatres.2008.01.013, 2008b.

Kumar, S., Reichle, R., Harrison, K., Peters-Lidard, C., Yatheendradas, S., and Santanello, J.: A comparison of methods for a priori bias correction in soil moisture data assimilation, Water Resources Research, in review, 2011.

Lange, H.: Are ecosystems dynamical systems?, International Journal of computing anticipatory systems, 3, 169–186, 1999.

Lawrence, D., Oleson, K., Flanner, M., Thornton, P., Swenson, S., Lawrence, P., Zeng, X., Yang, Z.-L., Levis, S., Sakaguchi, K., Bonan, G., and Slater, A.: Parameterization improvements and functional and structural advances in version 4 of the Community Land Model, J. Adv. Model. Earth Sys., 3, doi: 10.1029/2011MS000045, 2011.

Liu, Q., Reichle, R., Bindlish, R., Cosh, M., Crow, W., de Jeu, R., De Lannoy, G., Huffman, G., and Jackson, T.: The contributions of precipitation and soil moisture observations to the skill of soil moisture estimates in land data assimilation system, J. Hydrometeor., doi:10.1175/JHM-D-10.05000, 2011a.

Liu, Y., Brown, J., Demargne, J., and Seo, D.-J.: A wavelet-based approach to assessing timing errors in hydrologic predictions, Journal of Hydrology, 397, 210–224, 2011b.

Lohmann, D., Mitchell, K., Houser, P., Wood, E., Schaake, J., Robock, A., Cosgrove, B., Sheffield, J., Duan, Q., Luo, L., Higgins, W., Pinker, R., and Tarpley, J.: Streamflow and water balance intercomparisons of four land surface models in the North American Land Data Assimilation System project, J. Geophys. Res., 109, doi:10.1029/2003JD003517, 2004.

Mitchell, K., Lohmann, D., Houser, P., Wood, E., Schaake, J., Robock, A., Cosgrove, B., Sheffield, J., Duan, Q., Luo, L., Higgins, R., Pinker, R., Tarpley, J., Lettenmaier, D., Marshall, C., Entin, J., Pan, M., Shi, W., Koren, V., Meng, J., Ramsay, B., and Bailey, A.: The multi-institution North American Land Data Assimilation System (NLDAS): Utilizing multiple GCIP products and partners in a continental distributed hydrological modeling system, J. Geophys. Res., 109, doi:10.1029/2003JD003823, 2004.

Mo, K., Long, L., Xia, Y., Yang, S., Schemm, J., and Ek, M.: Drought indices based on the climate forecast system reanalysis and ensemble NLDAS, J. Hydrometeor., 12, 181–205, 2011.

Moore, B., Bertone, S., Mitchell, K., Rice, P., and Neill, R.: A worldwide near-real time diagnostic agrometeorological model, in: 20th AMS Conf. Ag and Forest Meteorology, pp. 7–11, 1990.

Mu, Q., Heinsch, F., Zhao, M., and Running, S.: Development of a Global evapotranspiration algorithm based on MODIS and global meteorology data, Remote Sensing of Environment, 111, 519–536, 2007.

Murphy, A. and Winkler, R.: Diagnostic verification of probability forecasts, International Journal of Forecasting, 7, 435–455, 1992.

Niu, G.-Y., Yang, Z.-L., Mitchell, K., Chen, F., Ek, M., Barlage, M., Kumar, A., Manning, K., Niyogi, D., Rosero, E., Tewari, M., and Xia, Y.-L.: The community Noah land surface model with multiparameterization options (Noah-MP): 1. Model description and evaluation with local-scale measurements, J. Geophys. Res., 116, doi:10.1029/2010JD015139, 2011.

Njoku, E., Jackson, T., Lakshmi, V., Chan, T., and Nghiem, S.: Soil moisture retrieval from AMSR-E, IEEE Transactions on Geoscience and Remote Sensing, 41, 215–229, 2003.

NRC: Assessment of Hydrologic and Hydrometeorological Operations and Services, Tech. rep., National Academy Press, 1996.

Owe, M., de Jeu, R., and Holmes, T.: Multi-sensor historical climatology of satellite-derived global land surface moisture, J. Geophys. Res., 13, doi:1029/2007JF000769, 2008.

Pachepsky, Y., Guber, A., Jacques, D., Simunek, J., van Genuchten, M., Nicholson, T., and Cady, R.: Information content and complexity of simulated soil water fluxes, Geoderma, pp. 253–266, 2006.

Pan, F., Pachepsky, Y., Andrey, G., and Hill, R.: Information and Complexity Measures Applied to Observed and Simulated Soil Moisture Time Series, Hydrological Sciences Journal, 56, 1027–1039, 2011.

Pan, M., Ming, J., Sheffield, J., Wood, E., Mitchell, K., Houser, P., Schaake, J., Robock, A., Lohmann, D., Cosgrove, B., Duan, Q., Luo, L., Higgins, R., Pinker, R., and Tarpley, J.: Snow process modeling in the North American Land Data Assimilation System (NLDAS): 2. Evaluation of model simulated snow water equivalent, J. Geophys. Res., 108, doi:10.1029/2003JD003994, 2003.

Peters-Lidard, C., Mocko, D., Garcia, M., Santanello, J., Tischler, M., and Moran, M. S.: Role of precipitation uncertainty in the estimation of hydrologic soil properties using remotely sensed soil moisture in a semiarid environment, Water Resources Research, 44, doi:10.1029/2007WR005884, 2008.

Peters-Lidard, C., Kumar, S., Mocko, D., and Tian, Y.: Estimating evapotranspiration with Land Data Assimilation Systems, Hydrological Processes, in print, 2011.

Peters-Lidard, C. D., Houser, P. R., Tian, Y., Kumar, S. V., Geiger, J., Olden, S., Lighty, L., Eastman, J. L., Doty, B., Dirmeyer, P., Adams, J., Mitchell, K., Wood, E., and Sheffield, J.: High-performance Earth System modeling with NASA/GSFC's Land Information System, Innovations in Systems and Software Engineering, 3, 157–165, 2007.

Pitman, A. and Henderson-Sellers, A.: Recent progress and results from the Project for the Intercomparison of Land surface Parameterization Schemes, J. Hydrol., 212-213, 128–135, 1998.

Pulliainen, J.: Mapping of snow water equivalent and snow depth in boreal and subarctic zones by assimilating space-borne microwave radiometer data and ground-based observations, Remote Sensing of Environment, 101, 257–269, 2006.

Randerson, J., Hoffman, F., Thornton, P., Mahowald, N., Lindsay, K., Lee, Y.-H., Nevison, C., Doney, S., Bonan, G., Stockli, R., Covey, C., Running, S., and Fung, I.: Systematic assessment of terrestrial biogeochemistry in coupled climate-carbon models, Global Change Biology, doi:10.1111/j.1365-2486.2009.01912.x, 2009.

Raupach, M., Rayner, P., Barrett, D., DeFries, R., Heimann, M., Ojima, D., Quegan, S., and Schmullius, C.: Model-data synthesis in terrestrial carbon observation: methods, data requirements and data uncertainty specifications, Global Change Biology, 11, 378–397, 2005.

Reichle, R. and Crow, W.: Comparison of adaptive filtering techniques for land surface data assimilation, Water Resources Research, 44, doi:10.1029/2007WR006883, 2008.

Reichle, R. and Koster, R.: Land data assimilation with the Ensemble Kalman Filter: Assessing model error parameters using innovations, in: Developments in Water Science - Computational Methods in Water Resources, edited by Hassanizadeh, S., Schotting, R., Gray, W., and Pinder, G., vol. 47, pp. 1387–1394, Elsevier, New York, 2002.

Reichle, R., Koster, R., Liu, P. and. Mahanama, S., Njoku, E., and Owe, M.: Comparison and assimilation of global soil moisture retrievals from the Advanced Microwave Scanning Radiometer for the Earth Observing System (AMSR-E) and the Scanning Multichannel Microwave Radiometer (SMMR), Journal of Geophysical Research-Atmospheres, 112, doi:10.1029/2006JD008033, 2007.

Reichle, R., Crow, W., and Keppenne, C.: An adaptive ensemble Kalman Filter for soil moisture data assimilation, Water Resources Research, 44, doi:10.1029/2007WR006357, 2008.

Robock, A., Luo, L., Wood, E., Wen, F., Mitchell, K., Houser, P., Schaake, J., Lohmann, D., Cosgrove, B.,

Sheffield, J., Duan, Q., Higgins, R., Pinker, R., Tarpley, J., Basara, J., and Crawford, K.: Evaluation of the North American Land Data Assimilation System over the southern Great Plains during the warm season, J. Geophys. Res., 108, 10.1029/2002JD003 245, 2003.

Rodell, M., Famiglietti, J., Chen, J., Seneviratne, S., Viterbo, P., Holl, S., and Wilson, C.: Basin scale estimates of evapotranspiration using GRACE and other observations, Geophys. Res. Lett., 31, doi: 10.1029/2004GL020873, 2004a.

Rodell, M., Houser, P. R., Jambor, U., Gottschalck, J., Mitchell, K., Meng., C.-J., Arsenault, K., Cosgrove, B., Radakovich, J., Bosilovich, M., Entin, J. K., Walker, J. P., Lohmann, D., and Toll, D.: The Global Land Data Assimilation System, Bull. Amer. Meteor. Soc., 85, 381–394, 2004b.

Rosenzweig, C., Tubiello, F., Goldberg, R., Mills, E., and Bloomfield, J.: Increased crop damage in the US from excess precipitation under climate change, Global Environmental Change, 12, 197–202, 2002.

Rossow, W. and Schiffer, R.: ISCCP cloud data products, Bull. Amer. Meteor. Soc., 72, 2–20, 1991.

Santanello, J., Peters-Lidard, C., Garcia, M., Mocko, D., Tischler, M., Moran, M., and Thoma, D.: Using remotely-sensed estimates of soil moisture to infer soil texture and hydraulic properties across a semi-arid watershed, Remote Sensing of Environment, 110, 79–97, 2007.

Santanello, J., Peters-Lidard, C., Kumar, S., Alonge, C., and Tao, W.-K.: A Modeling and Observational Framework for Diagnosing local land-atmosphere coupling on diurnal time scales, J. Hydrometeor., doi: 10.1175/2009JHM1066.1, 2009.

Santanello, J., Peters-Lidard, C., and Kumar, S.: Diagnosing the Sensitivity of Local Land-Atmosphere Coupling via the Soil Moisture-Boundary Layer Interaction, J. Hydrometeor., doi:10.1175/JHM-D-10-05014.1, 2011.

Selle, B. and Huwe, B.: Effective landscape modeling using CART and complexity measures, Geophysical Research Abstracts, 6, 00382, 2004.

Seyfried, M. and Wilcox, B.: Scale and the nature of spatial variability: Field examples having implications for hydrologic modeling, Water Resources Research, 31, 173–184, 1995.

Sheffield, J., Pan, M., Wood, E., Mitchell, K., Houser, P., Schaake, J., Robock, A., Lohmann, D., Cosgrove, B., Duan, Q., Luo, L., Higgins, R., Pinker, R., Tarpley, J., and Ramsay, B.: Snow process modeling in the North American Land Data Assimilation System (NLDAS): 1. Evaluation of model-simulated snow cover extent, J. Geophys. Res., 108, 10.1029/2002JD003 274, 2003.

Sivapalan, M. and Kalma, J.: Scale problems in hydrology: Contributions of the Robertson Workshop, Hydrological Processes, 9, 243–250, 1995.

Stanski, H., Wilson, L., and Burrows, W.: Survey of common verification methods in meteorology, Tech. report 8,WMO/TD 35, WMO World Weather Watch, 114 pp, 1989.

Trujillo, E., Ramirez, J., and Elder, K.: Scaling properties and spatial organization of snow depth fields in sub-alpine forest and alpine tundra, Hydrological Processes, doi:10.1002/hyp.7270, 2009.

van den Hurk, B., Best, M., Dirmeyer, P., Pitman, A., Polcher, J., and Santanello, J.: Over a decade of GLASS has accelerated land surface model development, Bull. Amer. Meteor. Soc., in print, 2011.

Wackerbauer, R., Witt, A., Atmanspacher, H., Kurths, J., and Scheingraber, H.: Comparative classification of complexity measures, Solitons Fractals, 4, 133–173, 1994.

Wealands, S., Grayson, R., and Walker, J.: Quantitative comparison of spatial fields for hydrological model as-

sessment - some promising approaches, Advances in Water Resources, 28, 15–32, doi:10.1016/j.advwatres.2004.10.001, 2005.

Williams, M., Richardson, A., Reichstein, M., Stoy, P., Peylin, P., Verbeeck, H., Carvalhais, N., Jung, M., Hollinger, D., Kattge, J., Leuning, R., Luo, Y., Tomelleri, E., Trudinger, C., and Wang, Y.-P.: Improving land surface models with FLUXNET data, Biogeosciences, 6, 1341–1359, 2009.

Wood, E., Sivapalan, M., Beven, K., and Band, L.: Effects of spatial variability and scale with implications to hydrologic modeling, J. Hydrol., 102, 29–47, 1988.

Wood, E., Sivapalan, M., and Beven, K.: Similarity and scale in catchment storm response, Reviews of Geophysics, 28, 1–18, doi:10.1029/RG028i001p00001, 1990.

Xia, Y., Mitchell, K., Ek, M., Cosgrove, B., Sheffield, J., Luo, L., Alonge, C., Wei, H., Meng, J., Livneh, B., Duan, Y., and Lohmann, D.: Continental-scale water and energy flux analysis and validation for North-American Land Data Assimilation System Project Phase-2, Part 2: Validation of model-simulated streamflow, Journal of Geophysical Research, in review, 2011a.

Xia, Y., Mitchell, K., Ek, M., Sheffield, J., Wood, E., Cosgrove, B., Luo, L., Alonge, C., Wei, H., Meng, J., Livneh, B., Lettenmaier, D.and Koren, V., , Duan, Y., K., M., and Fan, Y.: Continental-scale water and energy flux analysis and validation for the North American Land Data Assimilation System Project Phase 2 (NLDAS-2), Part 1: Comparison Analysis and Application of Model Products, Journal of Geophysical Research, in review, 2011b.

Xie, P. and Arkin, P.: Global precipitation: A 17-year monthly analysis based on gauge observations, satellite estimates, and numerical model outputs, Bull. Amer. Meteor. Soc., 78, 2539–2558, 1997.

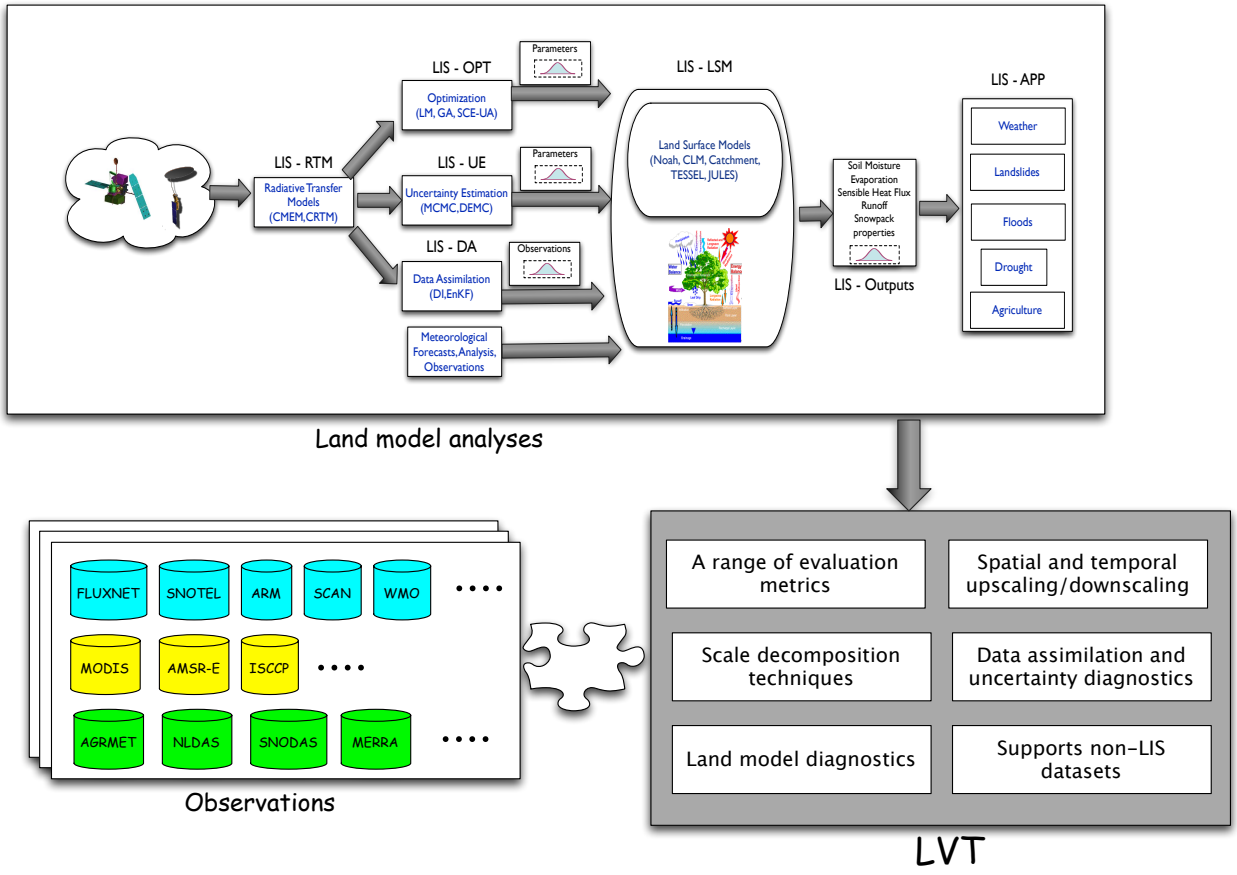


Fig. 1. Schematic of the Land surface Verification Toolkit and the association with the Land Information System (LIS). LVT supports the analysis of outputs from various LIS subsystems. LIS-DA represents the data assimilation subsystem, LIS-RTM represents the radiative transfer models within LIS, LIS-OPT represents the optimization subsystem, LIS-UE represents the uncertainty estimation subsystem, LIS-LSM represents the land surface models, and LIS-APP represents the various application models within LIS.

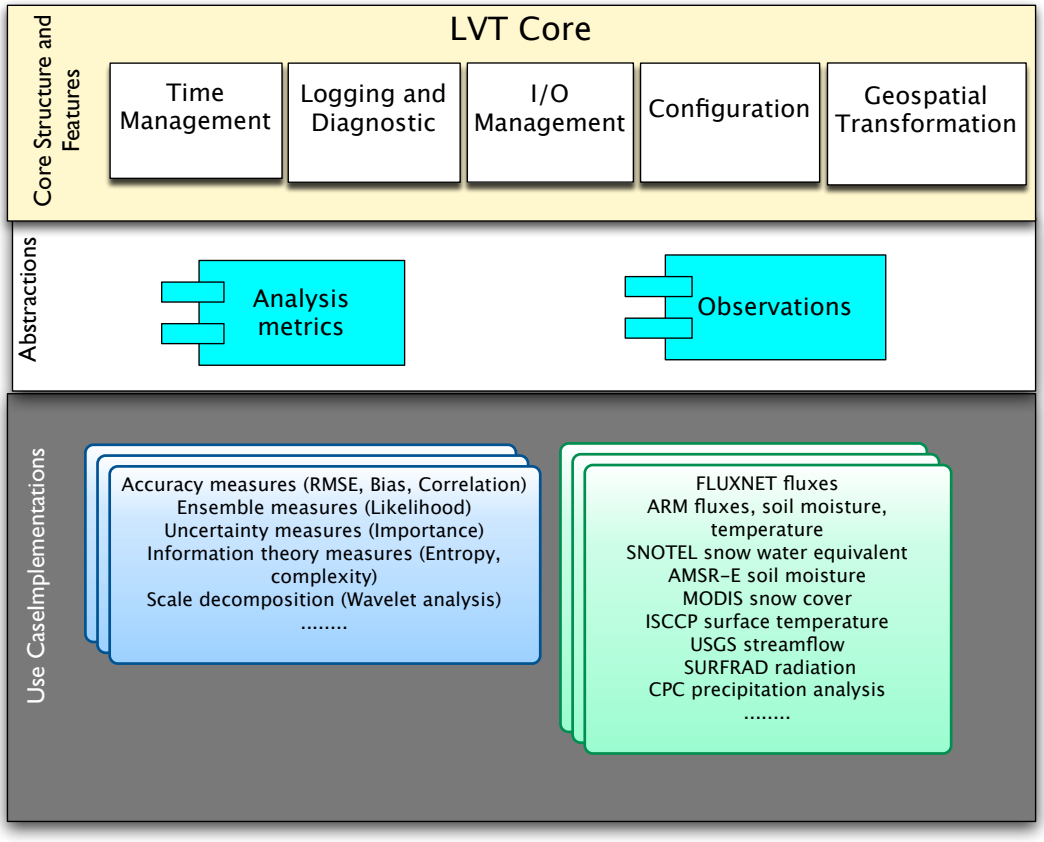


Fig. 2. Three-layer software architecture of Land surface Verification Toolkit (LVT)

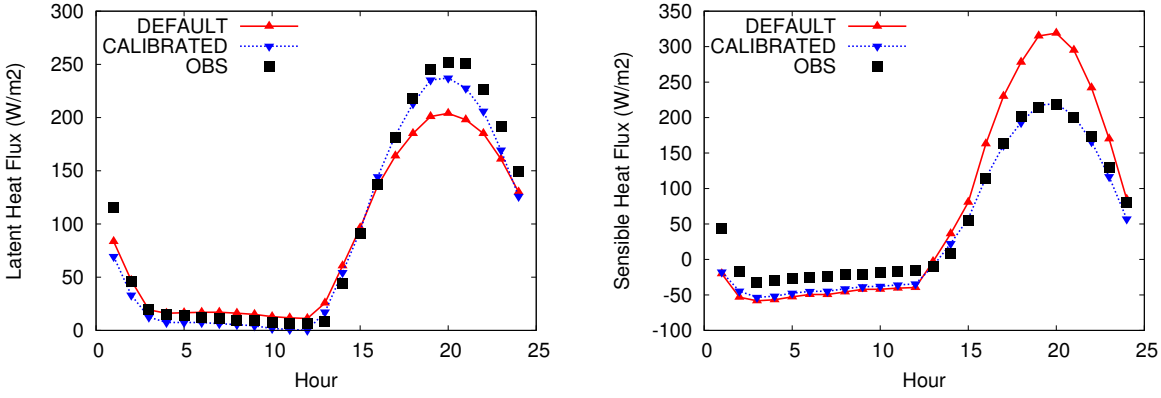


Fig. 3. Comparison of average diurnal cycles of latent (left column) and sensible heat (right column) fluxes from the uncoupled Noah (version 3.2) LSM simulations using the default model parameters (DEFAULT) and calibrated parameters (CALIBRATED) against the in-situ measurements (OBS) from 19 ARM-SGP stations.

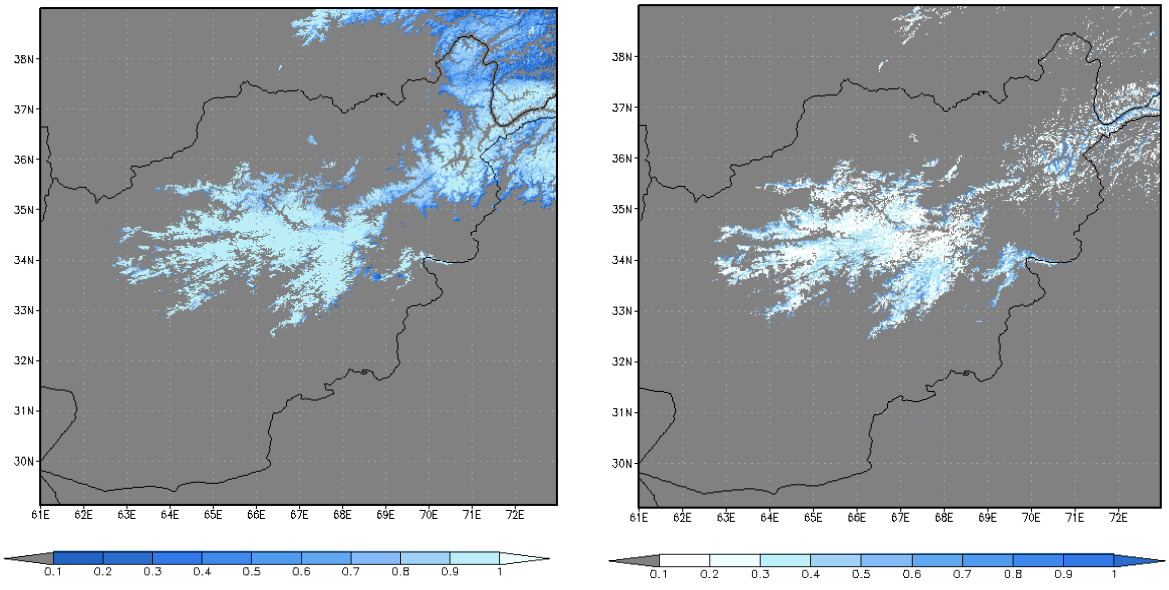


Fig. 4. Probability of Detection (left column) and False Alarm Ratio (right column) of the model simulated snow cover fields compared against the fractional MODIS snow cover product (MOD10A1).

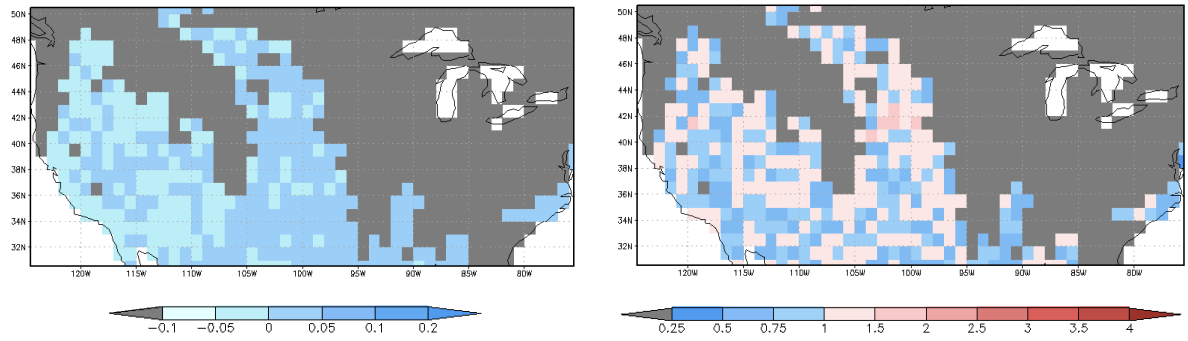
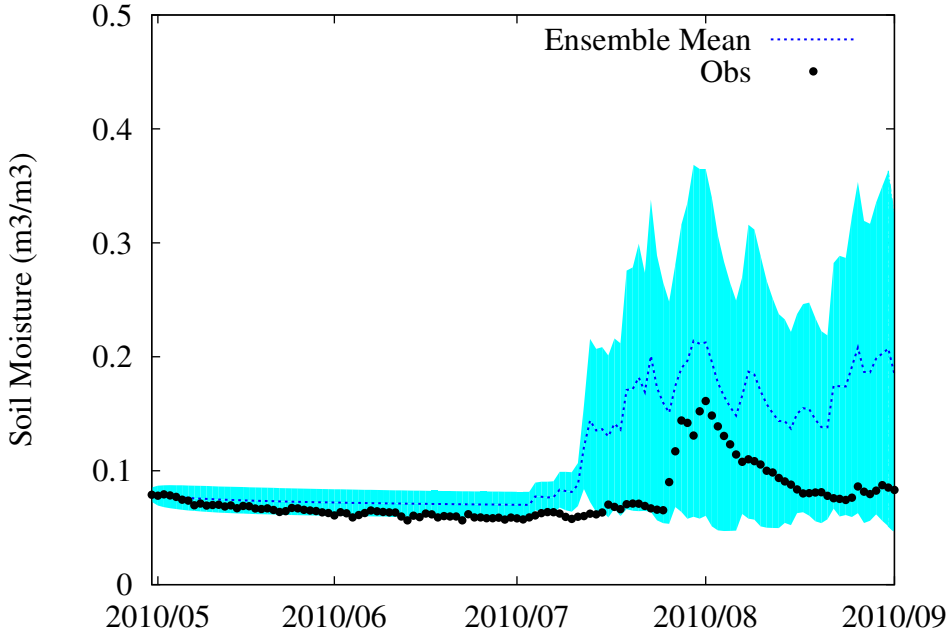
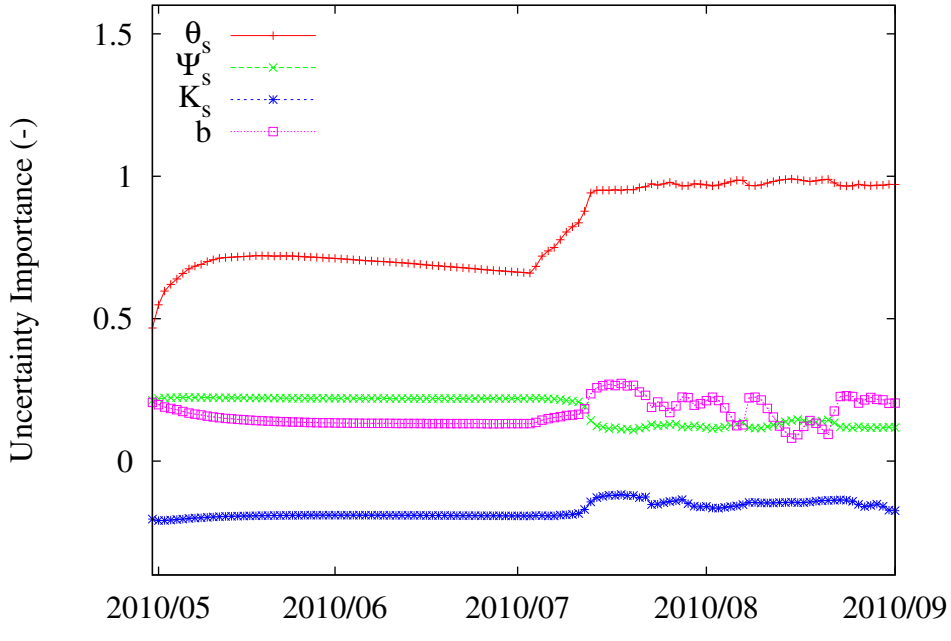


Fig. 5. Mean (left column) and variance (right column) of normalized innovations (dimensionless) of data assimilation diagnostics. The gray color represents grid cells excluded from the computations.



(a)



(b)

Fig. 6. (a) Comparison of ensemble soil moisture simulations against observations. The cyan shading indicates the ensemble spread, shown as $\pm$ 2 $\times$ ensemble standard deviation (b) The uncertainty importance of model parameters towards soil moisture uncertainty.

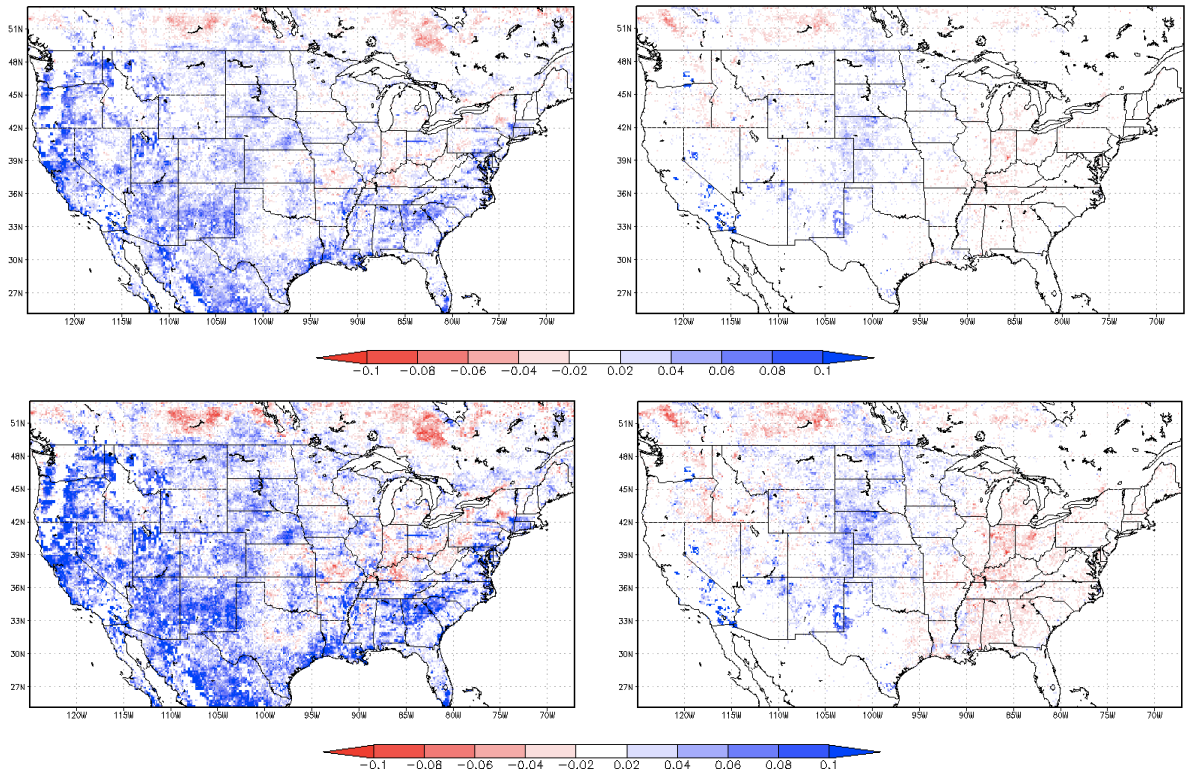


Fig. 7. Changes in Metric Entropy (top row) and Information gain (bottom row) from the assimilation of NASA AMSR-E (left column) and LPRM AMSR-E (right column) retrievals

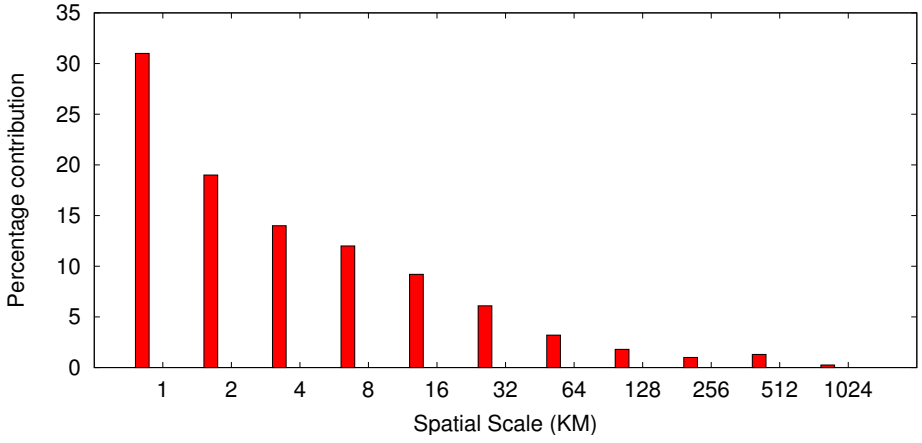


Fig. 8. Percentage contribution to the total improvement in snow covered area POD at different spatial scales, generated by a two dimensional discrete Haar wavelet analysis.

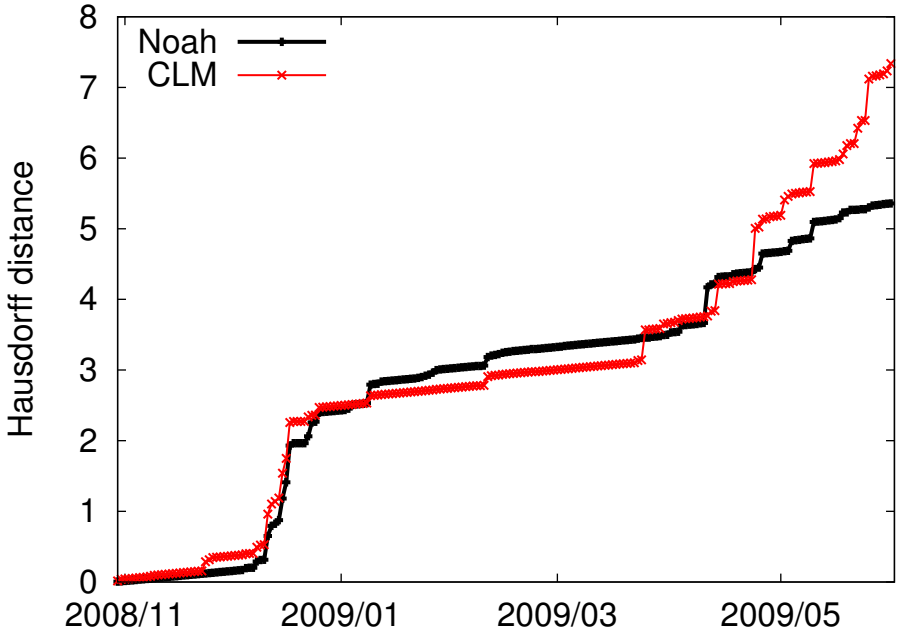


Fig. 9. Comparison of the cumulative Hausdorff distance measures of snow cover simulations from Noah and CLM

Table 1: List of datasets supported in LVT

Model/reanalysis outputs	
Agricultural Meteorology Model (AGRMET) from the Air Force Weather Agency (AFWA)	Water and energy fluxes, Soil moisture, Soil temperature, Snow conditions, Meteorology
NLDAS model outputs (Mitchell et al. (2004))	Water and energy fluxes Soil moisture, Soil temperature, Snow conditions, Meteorology
GLDAS model outputs (Rodell et al. (2004b))	Water and energy fluxes, Soil moisture, Soil temperature, Snow conditions, Meteorology
Canadian Meteorological Center (CMC) snow depth analysis (Brown and Brasnett (2010))	Snow depth
Snow Data Assimilation System (SNODAS; Barrett (2003))	Snow depth, Snow water equivalent
In-situ measurements	
AMMA (database.amma-international.org/)	Water and energy fluxes, soil moisture, soil temperature
Atmospheric Radiation Measurement (ARM) (www.arm.gov)	Water and energy fluxes, Soil moisture, soil temperature, Meteorology
Ameriflux (public.ornl.gov/ameriflux/)	Water and energy fluxes
Coordinated Energy and water cycle Observations Project (CEOP) (www.ceop.net/)	Water and energy fluxes, Soil moisture, soil temperature, Meteorology
National Weather Service	Snow depth, Precipitation,

Dataset	Measurement variables
Cooperative Observer Program (COOP) (www.nws.noaa.gov/om/coop/)	Land surface temperature
NOAA CPC unified (Higgins et al. (1996))	Precipitation
Gridded FLUXNET (Jung et al. (2009))	Water and energy fluxes
Finnish Meteorological Institute (FMI/SYKE; www.environment.fi/syke)	Snow water equivalent
Global Summary of the Day (GSOD)	Snow depth
International Soil Moisture Network (www.ipf.tuwien.ac.at/insitu/)	Soil moisture
Soil Climate Analysis Network (SCAN; www.wcc.nrcs.usda.gov/scan/)	Soil moisture Soil temperature
WMO synoptic observations	Snow depth
NRCS SNOwpack TELemetry network (SNOTEL;www.wcc.nrcs.usda.gov/snow/)	Snow water equivalent
Surface Radiation Network (SURFRAD) (www.srrb.noaa.gov/surfrad/)	Downwelling shortwave, Downwelling longwave
Southwest Watershed Research Center (SWRC; www.tucson.ars.ag.gov/dap/)	Soil moisture, Soil temperature
USGS water data (waterdata.usgs.gov/nwis)	Streamflow

Dataset	Measurement variables
AMSR-E radiances (mrain.atmos.colostate.edu/LEVEL1C/)	Brightness temperature for different channels
Satellite and remote sensing data	
AFWA NASA Snow Algorithm (ANSA; Foster et al. (2011))	Snow cover, Snow depth, Snow water equivalent
GlobSnow (Pulliainen (2006)) (www.globsnow.info/)	Snow cover, Snow water equivalent
International Satellite Cloud Climatology Project (ISCCP; Rossow and Schiffer (1991)) (isccp.nasa.gov)	Land surface temperature
MODIS/Terra Snow cover 500m (MOD10A1; Hall et al. (2006))	Snow cover
MODIS Evapotranspiration product (MOD16; Mu et al. (2007))	Evapotranspiration
NASA Level-3, soil moisture retrieval from AMSR-E (AE_Land3) Njoku et al. (2003)	Soil moisture
Land Parameter Retrieval Model (LPRM) from NASA GSFC and VU Amsterdam (Owe et al. (2008))	Soil moisture

Table 2. The range of analysis metric types and implementations supported in LVT

Metric class	Supported Implementations
Standard measures	RMSE, Anomaly RMSE, unbiased RMSE (ubRMSE), Correlation, Anomaly correlation, Mean absolute error (MAE), Bias, Probability of "yes" detection (PODy), False alarm ratio (FAR) Probability of "no" detection (PODn), Accuracy measure (ACC), Probability of false detection (POFD), Critical success index (CSI), Equitable threat score (ETS), Frequency bias (FBIAS), Nash sutcliffe efficiency (NSE)
Ensemble metrics	Mean, Standard deviation, Likelihood
Uncertainty metrics	Uncertainty importance
Information theoretic	Metric entropy, Information gain, Effective complexity, Fluctuation complexity
Data assimilation metrics	Mean, variance, lag correlation of innovation distributions
Spatial similarity metrics	Spatial area, Hausdorff distance
Scale decomposition	Discrete wavelet transforms